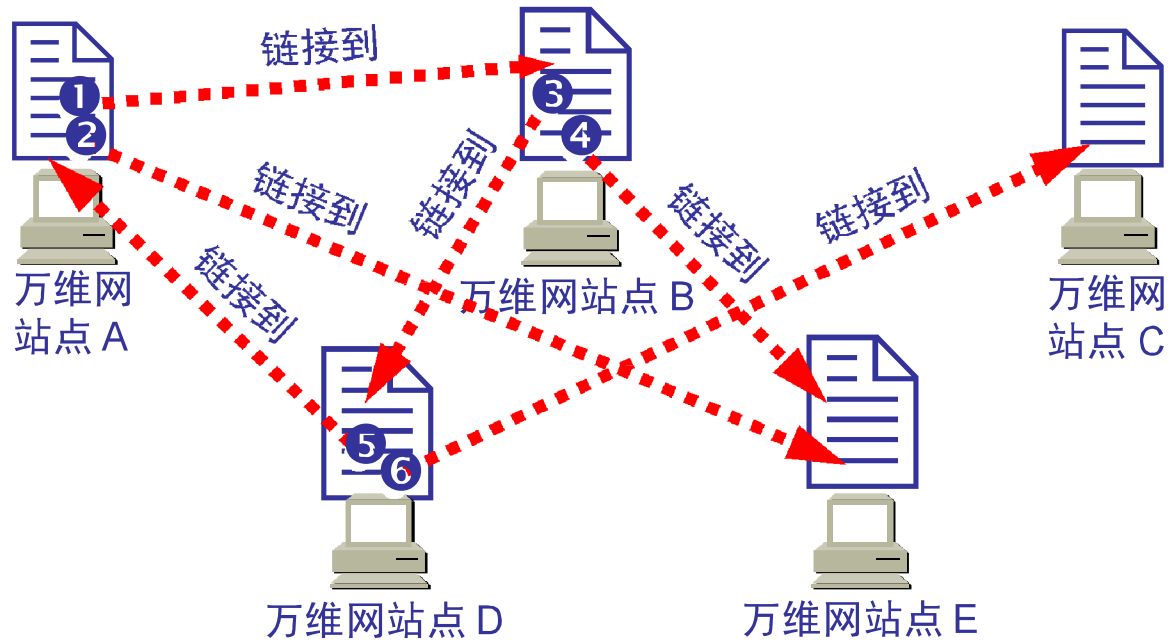


6.3 万维网 WWW

6.3.1 万维网概述

- **万维网** WWW (World Wide Web)并非某种特殊的计算机网络。
- 万维网是一个大规模的、联机式的信息储藏所。
- 万维网用链接的方法能非常方便地从因特网上的一个站点访问另一个站点，从而主动地按需获取丰富的信息。
- 这种访问方式称为“**链接**”。

万维网提供分布式服务





超媒体与超文本

- 万维网是**分布式超媒体**(hypermedia)系统，它是**超文本**(hypertext)系统的扩充。
- 一个超文本由多个信息源链接成。利用一个链接可使用户找到另一个文档。这些文档可以位于世界上任何一个接在因特网上的超文本系统中。超文本是万维网的基础。
- 超媒体与超文本的区别是文档内容不同。超文本文档仅包含文本信息，而超媒体文档还包含其他表示方式的信息，如图形、图像、声音、动画，甚至活动视频图像。



万维网的工作方式

- 万维网以客户服务器方式工作。
- **浏览器**就是在用户计算机上的万维网**客户程序**。万维网文档所驻留的计算机则运行**服务器程序**，因此这个计算机也称为**万维网服务器**。
- 客户程序向服务器程序发出请求，服务器程序向客户程序送回客户所要的万维网文档。
- 在一个客户程序主窗口上显示出的万维网文档称为**页面**(page)。



万维网必须解决的问题

- (1) 怎样标志分布在整个因特网上的万维网文档?
- 使用**统一资源定位符** URL (Uniform Resource Locator)来标志万维网上的各种文档。
 - 使每一个文档在整个因特网的范围内具有唯一的标识符 URL。



万维网必须解决的问题

(2) 用何协议实现Web页面的传输？

- 在万维网客户程序与万维网服务器程序之间进行交互所使用的协议，是**超文本传送协议** HTTP (HyperText Transfer Protocol)。
- HTTP 是一个应用层协议，它使用 TCP 连接进行可靠的传送。



万维网必须解决的问题

- (3) 如何编写万维网文档使它们能在因特网上的各种计算机上显示出来，如何在文档中嵌入超链？
- **超文本标记语言** HTML (HyperText Markup Language)使得万维网页面的设计者可以很方便地用一个超链从本页面的某处链接到因特网上的任何一个万维网页面，并且能够在自己的计算机屏幕上将这些页面显示出来。



万维网必须解决的问题

- (4) 怎样使用户能够很方便地找到所需的信息？
- 为了在万维网上方便地查找信息，用户可使用各种的搜索工具（即搜索引擎）。

6.3.2 统一资源定位符 URL

1. URL的格式

- 统一资源定位符 URL 是对可以从因特网上得到的资源的位置和访问方法的一种简洁的表示。
- URL 给资源的位置提供一种抽象的识别方法，并用这种方法给资源定位。
- 只要能够对资源定位，系统就可以对资源进行各种操作，如存取、更新、替换和查找其属性。
- URL 相当于一个文件名在网络范围的扩展。因此 URL 是与因特网相连的机器上的任何可访问对象的一个指针。



URL 的一般形式

- 由以冒号隔开的两大部分组成，并且在 URL 中的字符对大写或小写没有要求。
- URL 的一般形式是：

<协议>://<主机>:<端口>/<路径>

ftp —— 文件传送协议 FTP

http —— 超文本传送协议 HTTP

News —— USENET 新闻

URL 的一般形式（续）

- 由以冒号隔开的两大部分组成，并且在 URL 中的字符对大写或小写没有要求。
- URL 的一般形式是：

<协议>://<主机>:<端口>/<路径>

<主机> 是存放资源的主机
在因特网中的域名

URL 的一般形式（续）

- 由以冒号隔开的两大部分组成，并且在 URL 中的字符对大写或小写没有要求。
- URL 的一般形式是：

<协议>://<主机>:<端口>/<路径>

有时可省略



2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

`http` //<主机>:<端口>/<路径>

↑
这表示使用 HTTP 协议



2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

http://<主机>:<端口>/<路径>

冒号和两个斜线是规定的格式



2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

http://<主机><端口>/<路径>

这里写主机的域名



2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

http://<主机>:<端口>/<路径>

HTTP 的默认端口号是 80，通常可省略



2. 使用 HTTP 的 URL

- 使用 HTTP 的 URL 的一般形式

http://<主机>:<端口>/<路径>

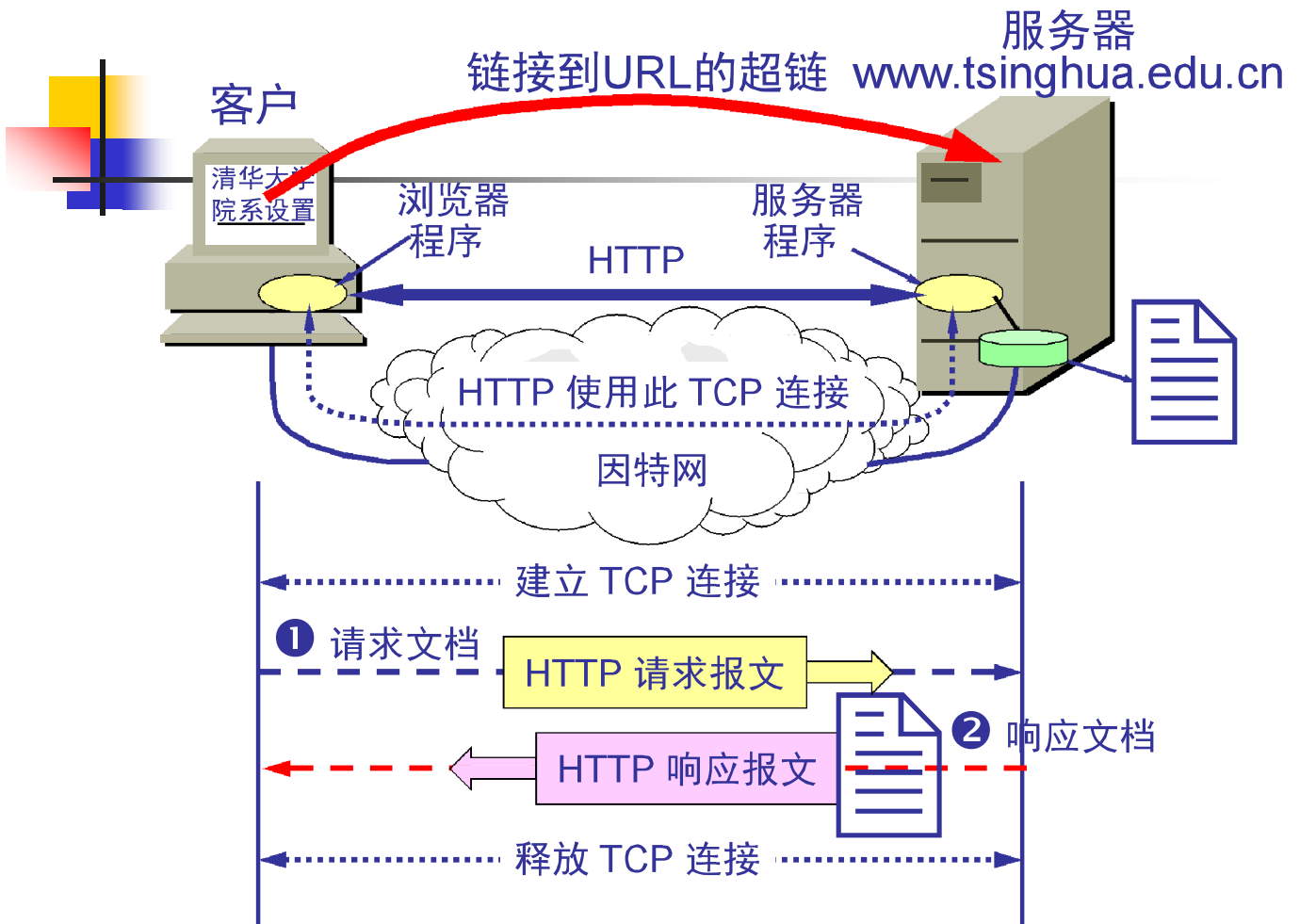
若再省略文件的<路径>项，则 URL 就指到因特网上的某个**主页**(home page)。

6.3.3 超文本传送协议 HTTP

1. HTTP 的操作过程

- HTTP协议定义了浏览器（即万维网客户进程）怎样向万维网服务器请求万维网文档，以及万维网服务器怎样把万维网文档传送给浏览器。
- HTTP使用的运输层协议是TCP，默认端口号是80。

万维网的工作过程



用户点击鼠标后所发生的事件

- (1) 浏览器分析超链指向页面的 URL。
- (2) 浏览器向 DNS 请求解析 `www.tsinghua.edu.cn` 的 IP 地址。
- (3) 域名系统 DNS 解析出清华大学服务器的 IP 地址。
- (4) 浏览器与服务器建立 TCP 连接
- (5) 浏览器发出取文件命令：
GET /chn/yxszt/index.htm。
- (6) 服务器给出响应，把文件 `index.htm` 发给浏览器。
- (7) TCP 连接释放。
- (8) 浏览器显示“清华大学院系设置”文件 `index.htm` 中的所有文本。



HTTP 的主要特点

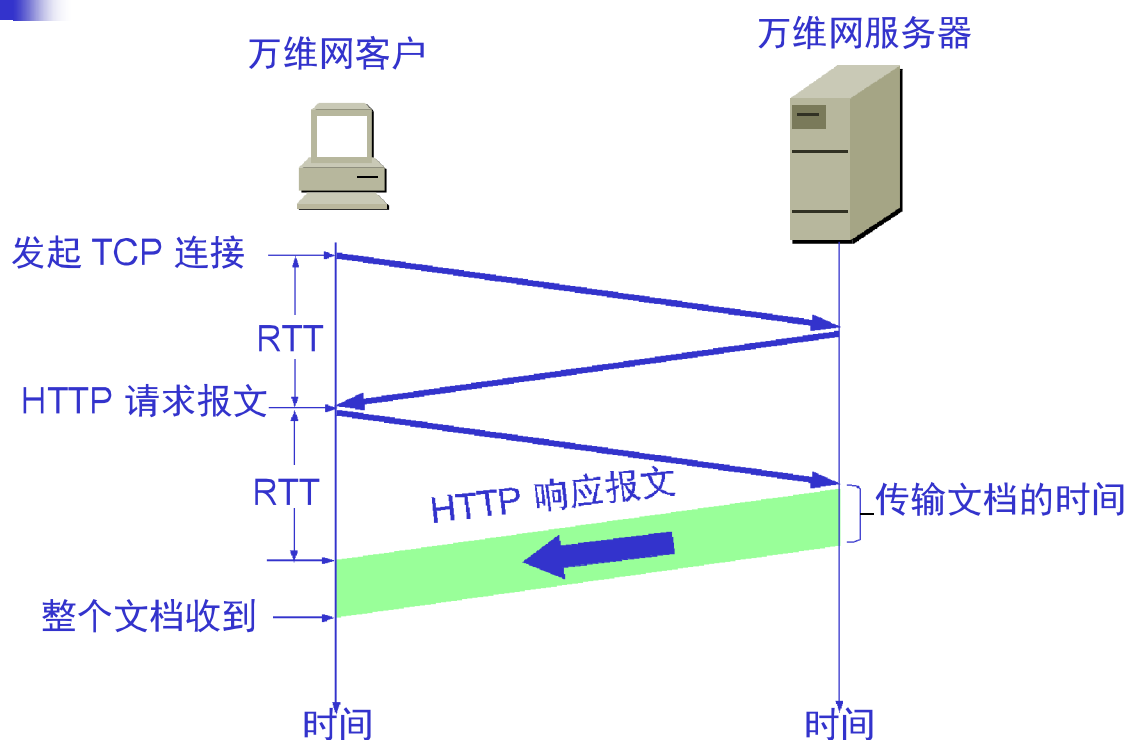
- HTTP 1.0 协议是**无状态的**(stateless)。HTTP不要求服务器保留客户的任何状态信息。
- HTTP 协议本身也是无连接的，虽然它使用了面向连接的 TCP 向上提供的服务。



2. 非持续连接与持续连接

- HTTP/1.0 协议使用非持续连接。
- 客户（浏览器）每发送一个请求，Web服务器在发送响应后就关闭这条连接
- 要向同服务器发送下一个请求需再建立TCP连接
- 一个Web网页除了一个基本的HTML文档外，可能还包括多个要在页面上显示或表现的引用对象（图片、声音等）
- 请求一个网页可能要建立多次连接，**效率太低**

请求一个万维网文档所需的时间





持续连接 (persistent connection)

- HTTP/1.1 协议使用持续连接。
- 万维网服务器在发送响应后仍然在一段时间内保持这条连接，使同一个客户（浏览器）和该服务器可以继续在这条连接上传送后续的 HTTP 请求报文和响应报文。
- 这并不局限于传送同一个页面上链接的文档，而是只要这些文档都在同一个服务器上就行。
- 目前一些流行的浏览器的默认设置就是使用 HTTP/1.1。



持续连接的两种工作方式

- 非流水线方式：客户在收到前一个响应后才能发出下一个请求。这比非持续连接的两倍 RTT 的开销节省了建立 TCP 连接所需的一个 RTT 时间。但服务器在发送完一个对象后，其 TCP 连接就处于空闲状态，浪费了服务器资源。
- 流水线方式：客户在收到 HTTP 的响应报文之前就能够接着发送新的请求报文。一个接一个的请求报文到达服务器后，服务器就可连续发回响应报文。使用流水线方式时，客户访问所有的对象只需花费一个 RTT 时间，使 TCP 连接中的空闲时间减少，提高了下载文档效率。

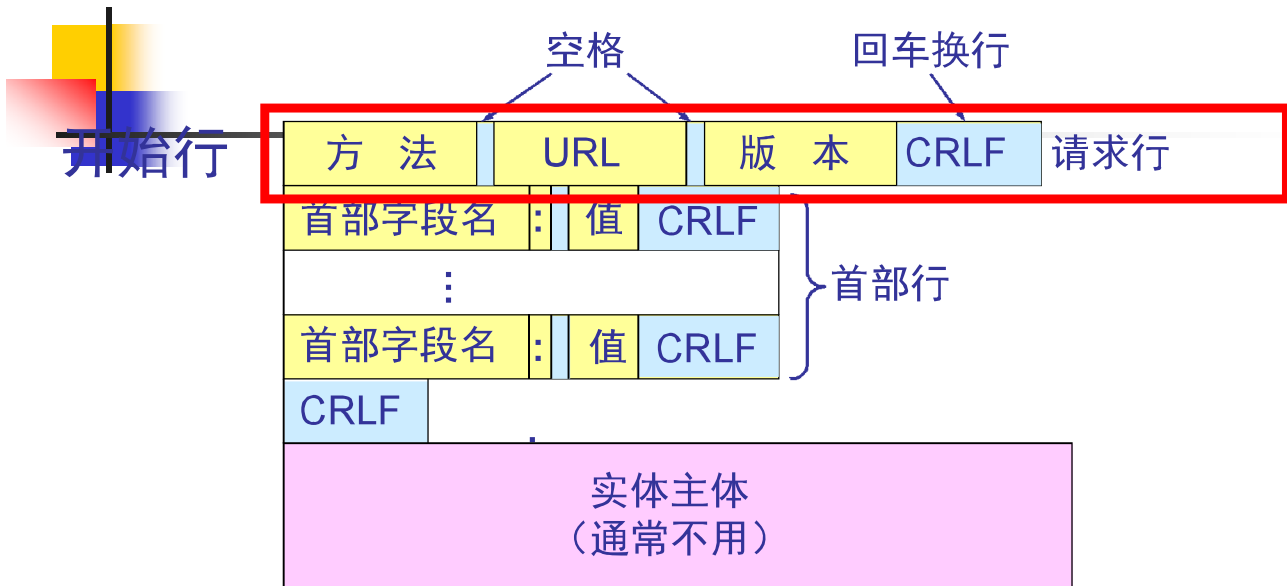


3. HTTP 的报文结构

HTTP 有两类报文：

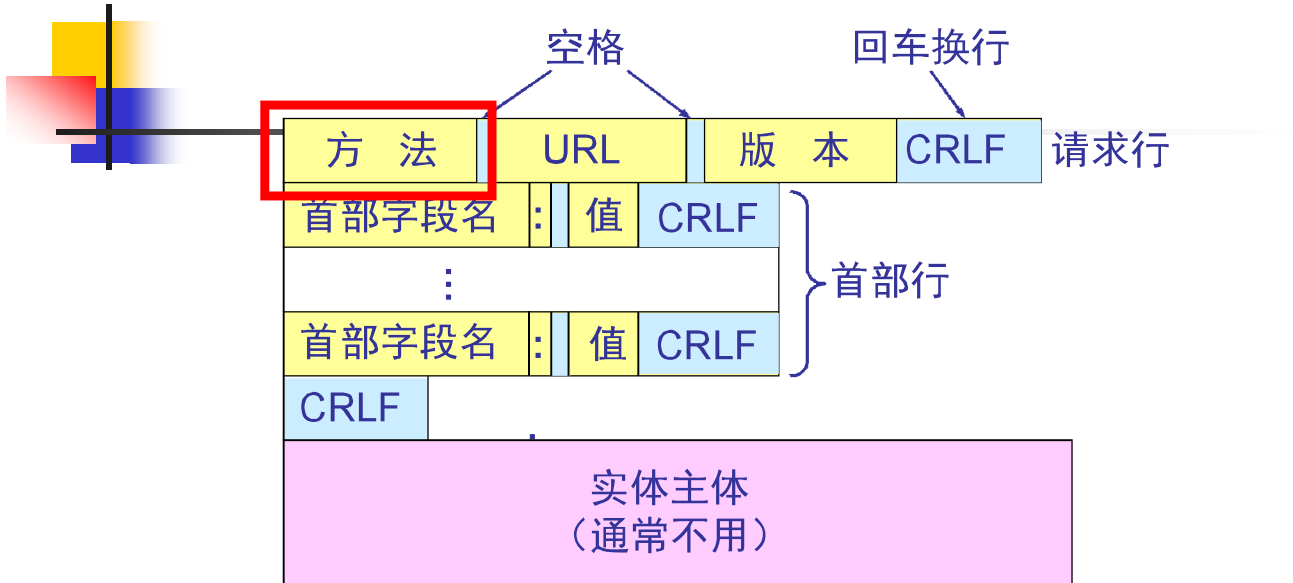
- 请求报文——从客户向服务器发送请求报文。
- 响应报文——从服务器到客户的回答。
- 由于 HTTP 是面向正文的(text-oriented)，因此报文中的每一个字段都是一些 ASCII 码串，因而每个字段的长度都是不确定的。

HTTP 的报文结构（请求报文）



报文由三个部分组成，即**开始行**、**首部行**和**实体主体**。在请求报文中，开始行就是请求行。

HTTP 的报文结构（请求报文）



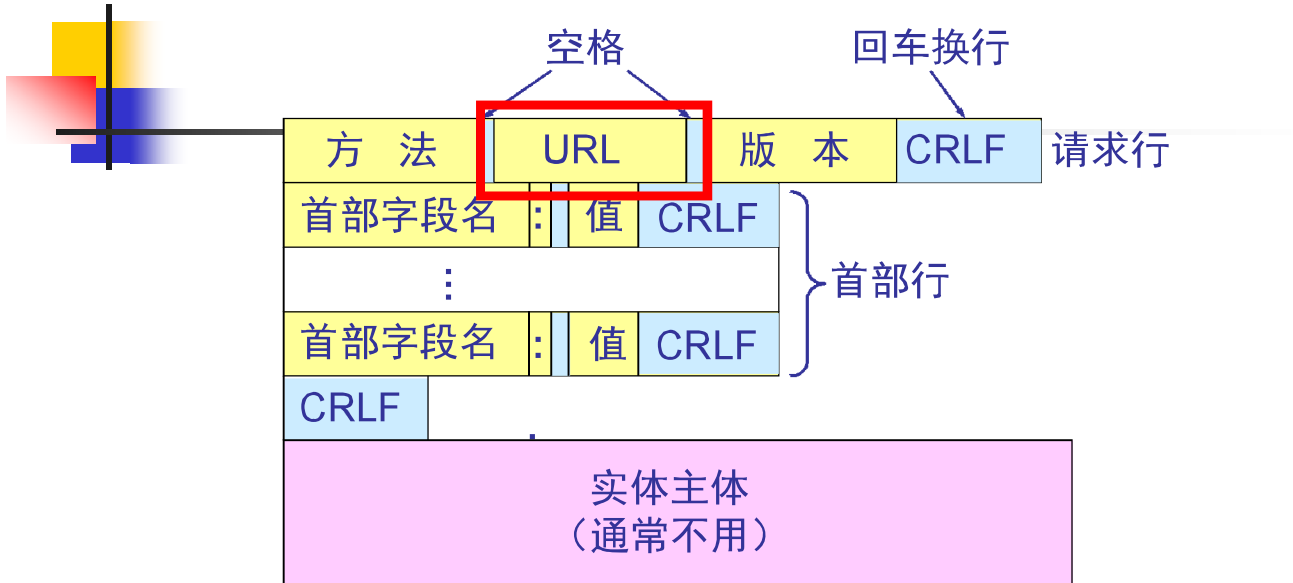
“**方法**”是面向对象技术中使用的专门名词。所谓“**方法**”就是**对所请求的对象进行的操作**，因此这些方法实际上也就是一些**命令**。因此，请求报文的类型是由它所采用的方法决定的。



HTTP 请求报文的一些方法

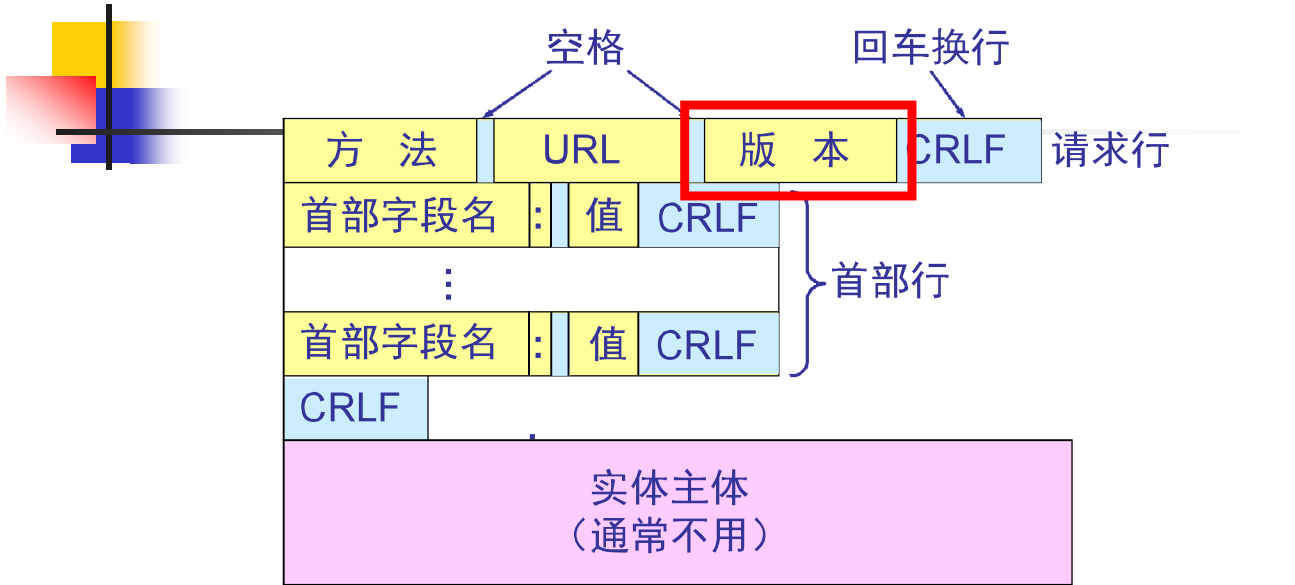
方法（操作）	意义
OPTION	请求一些选项的信息
GET	请求读取由 URL 所标志的信息
HEAD	请求读取由 URL 所标志的信息的首部
POST	给服务器添加信息（例如，注释）
PUT	在指明的 URL 下存储一个文档
DELETE	删除指明的 URL 所标志的资源
TRACE	用来进行环回测试的请求报文
CONNECT	用于代理服务器

HTTP 的报文结构（请求报文）



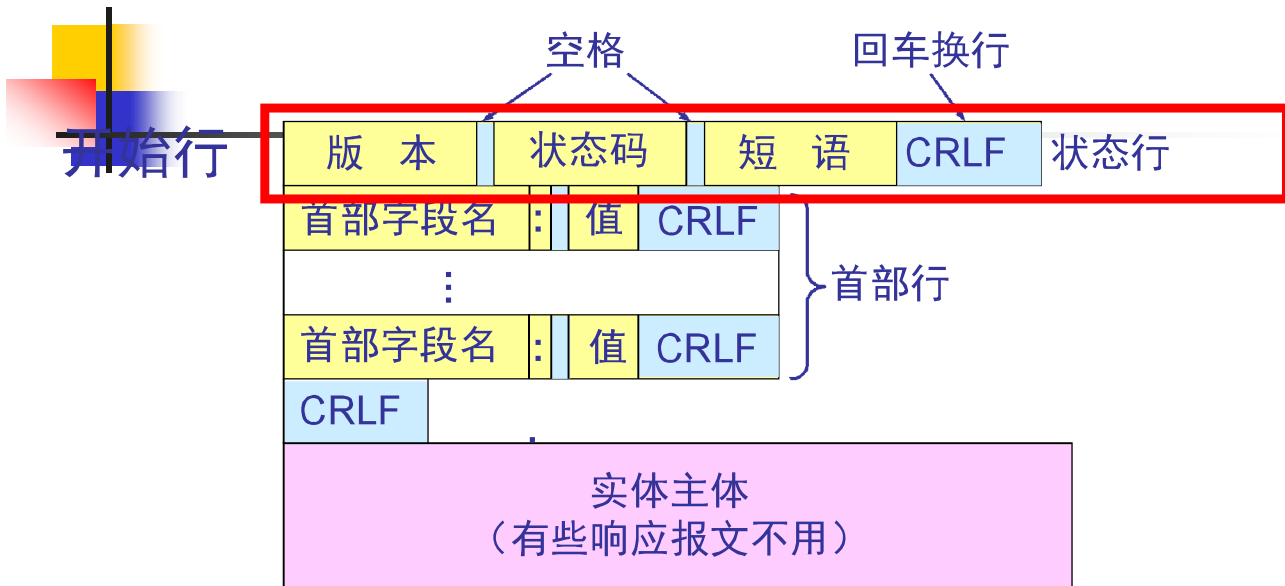
“URL”是所请求的资源的 URL。

HTTP 的报文结构（请求报文）



“版本” 是 HTTP 的版本。

HTTP 的报文结构（响应报文）



响应报文的开始行是**状态行**。
状态行包括三项内容，即 **HTTP 的版本**，**状态码**，
以及解释状态码的**简单短语**。



状态码都是三位数字

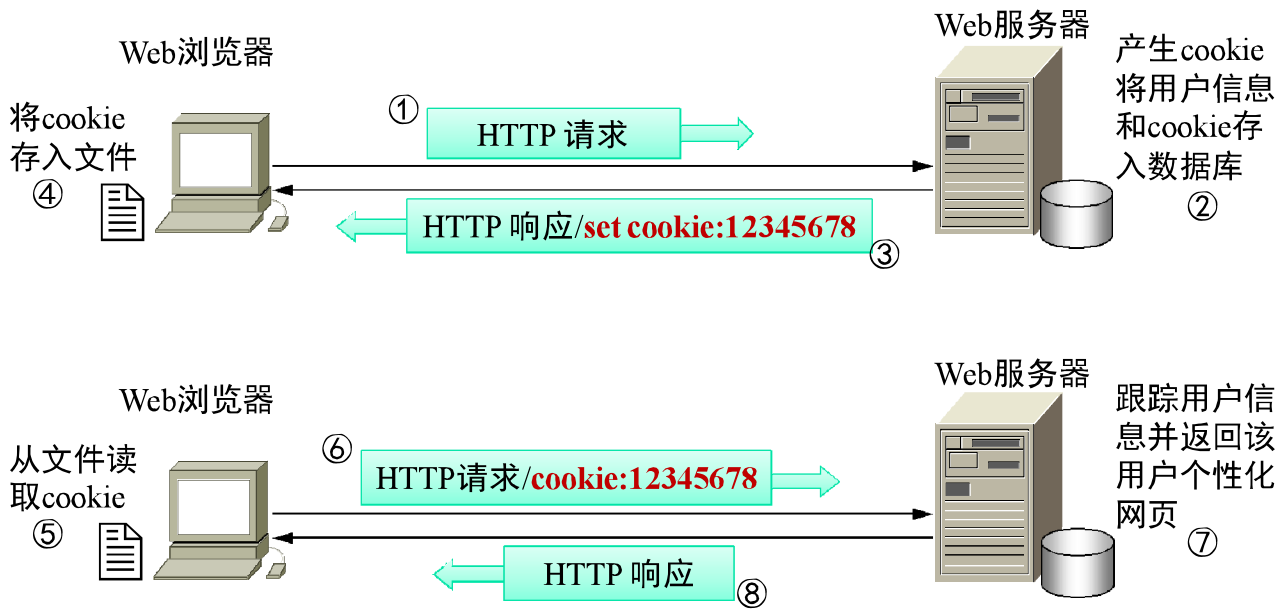
- 1xx 表示通知信息的，如请求收到了或正在进行处理。
- 2xx 表示成功，如接受或知道了。
- 3xx 表示重定向，表示要完成请求还必须采取进一步的行动。
- 4xx 表示客户的差错，如请求中有错误的语法或不能完成。
- 5xx 表示服务器的差错，如服务器失效无法完成请求。



4. 在服务器上存放用户的信息

- 万维网站点使用 Cookie 来跟踪用户。
- Cookie 表示在 HTTP 服务器和客户之间传递的状态信息。
- 使用 Cookie 的网站服务器为用户产生一个唯一的识别码。利用此识别码，网站就能够跟踪该用户在该网站的活动。

4. 在服务器上存放用户的信息



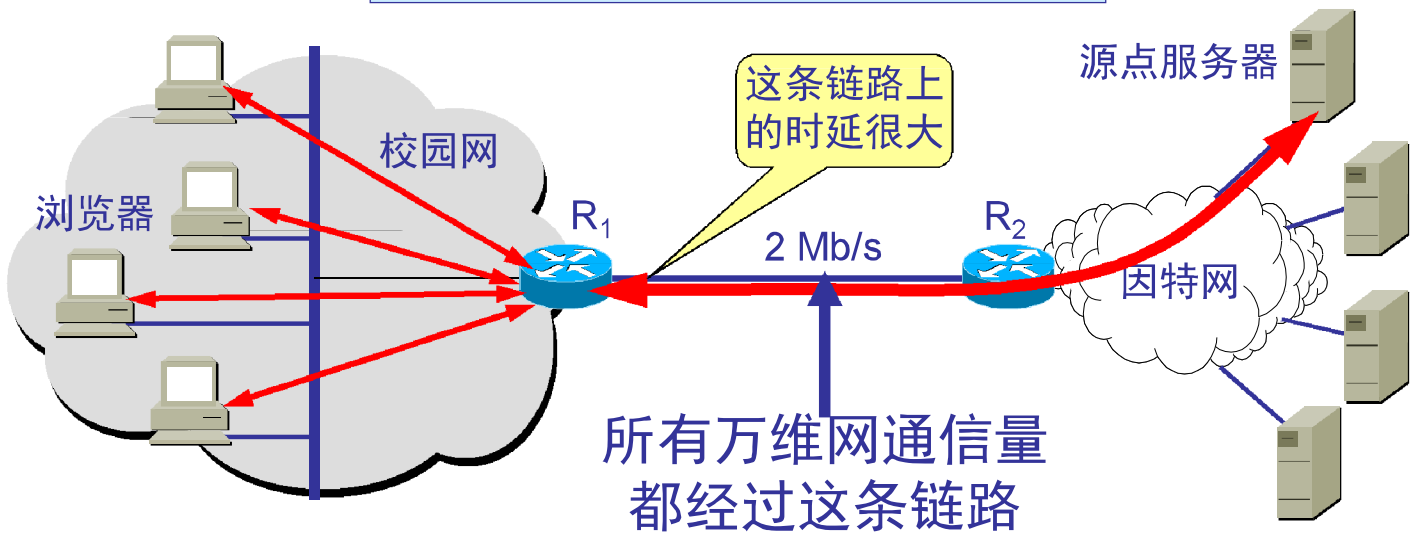


5. 代理服务器 (proxy server)

- **代理服务器**(proxy server)又称为万维网高速缓存(Web cache), 它代表浏览器发出 HTTP 请求。
- 万维网高速缓存把最近的一些请求和响应暂存在本地磁盘中。
- 当与暂时存放的请求相同的新请求到达时, 万维网高速缓存就把暂存的响应发送出去, 而不需要按 URL 的地址再去因特网访问该资源。

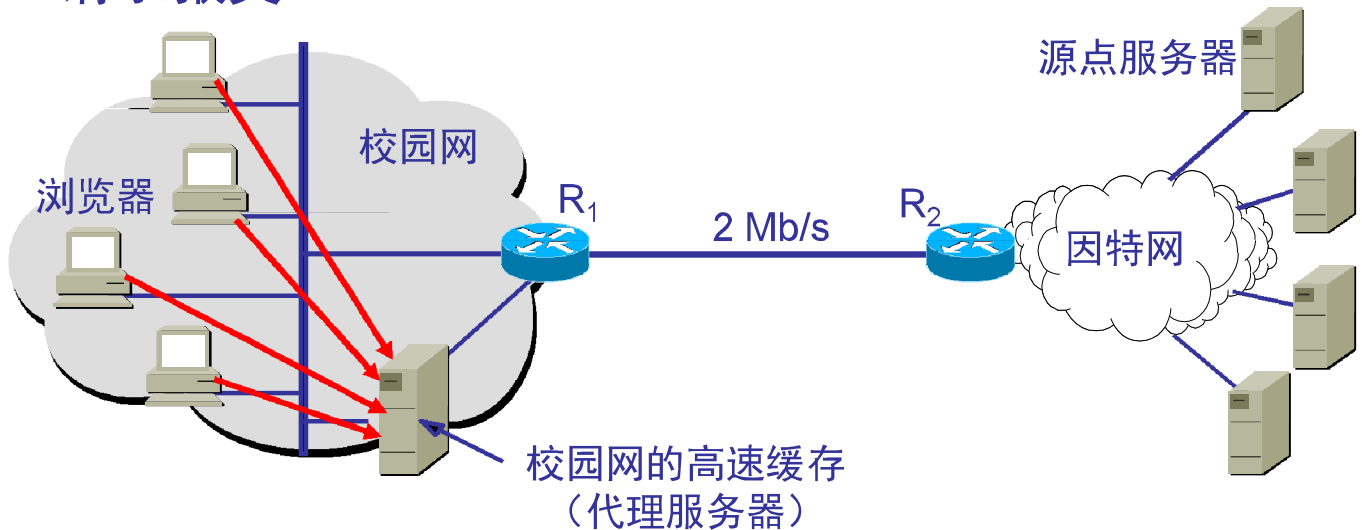
使用高速缓存可减少 访问因特网服务器的时延

没有使用高速缓存的情况



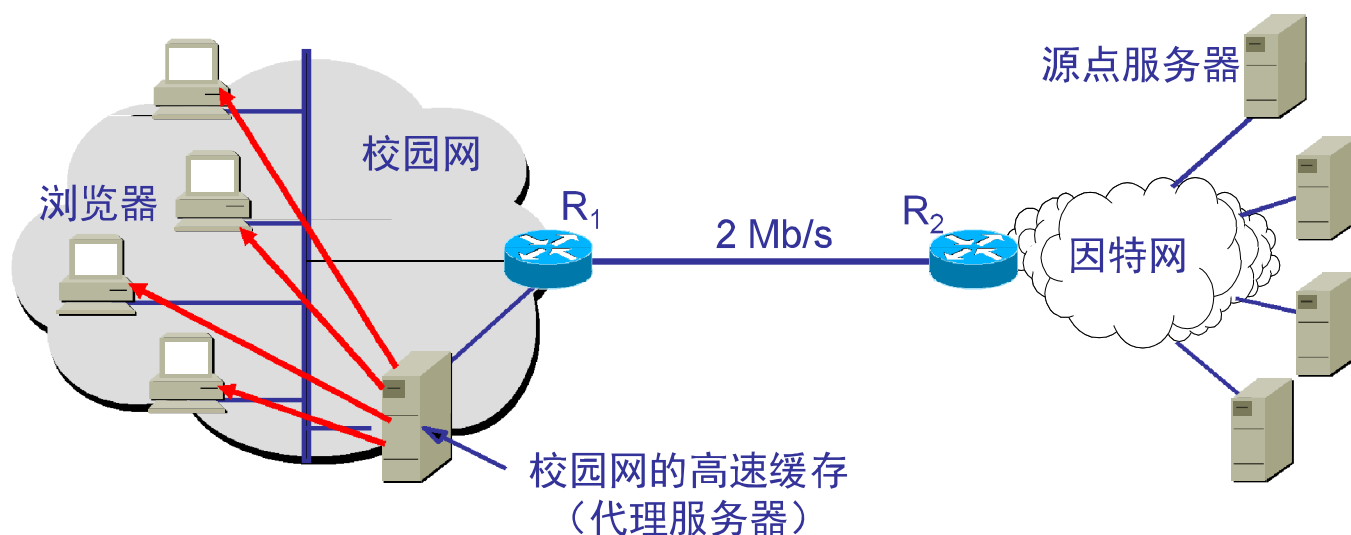
使用高速缓存的情况

(1) 浏览器访问因特网的服务器时，要先与校园网的高速缓存建立 TCP 连接，并向高速缓存发出 HTTP 请求报文



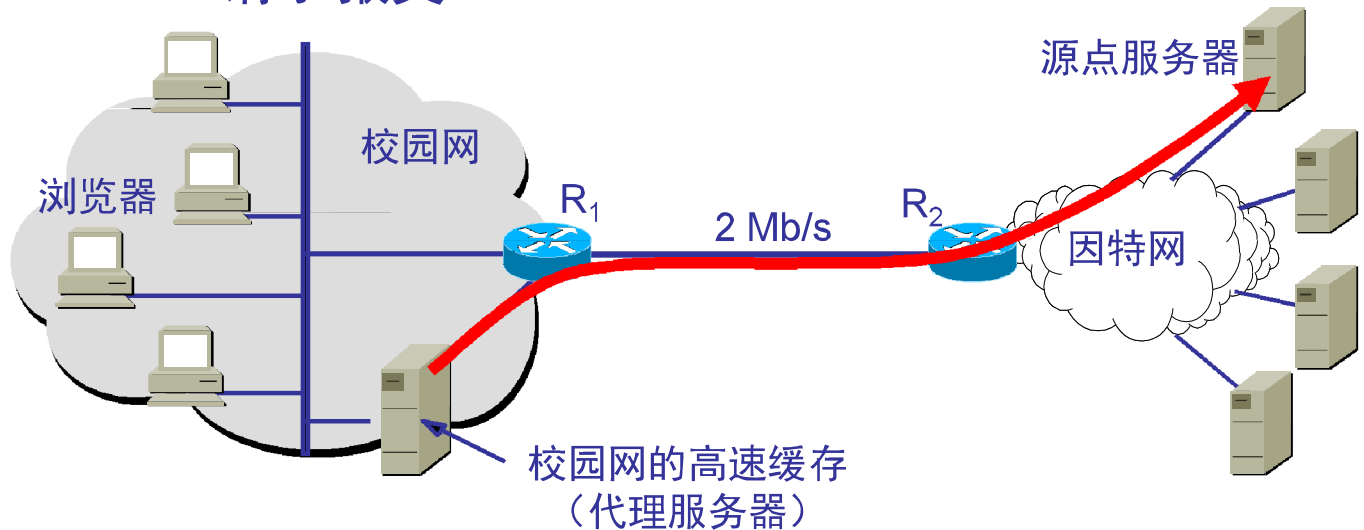
使用高速缓存的情况

(2) 若高速缓存已经存放了所请求的对象，则将此对象放入 HTTP 响应报文中返回给浏览器。



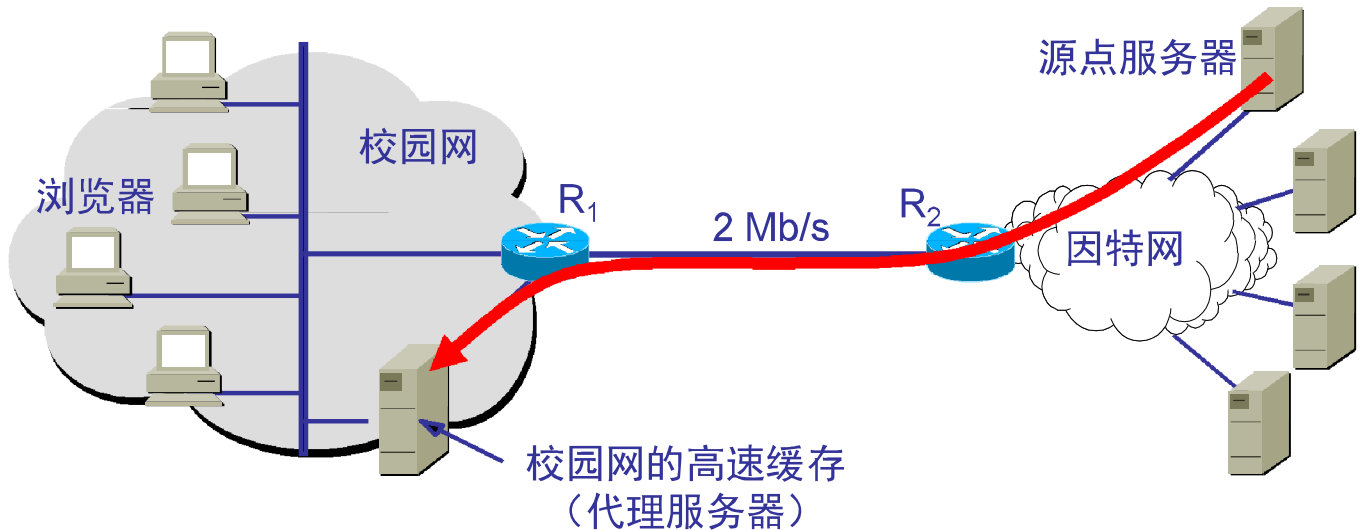
使用高速缓存的情况

(3) 否则，高速缓存就代表发出请求的用户浏览器，与因特网上的源点服务器建立 TCP 连接，并发送 HTTP 请求报文。



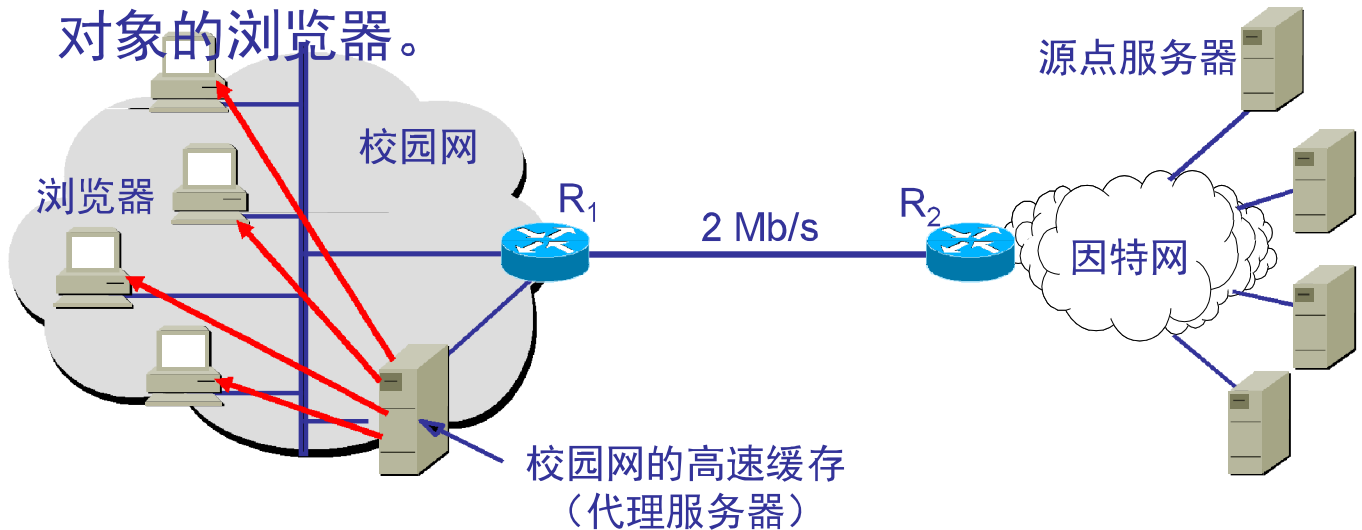
使用高速缓存的情况

(4) 源点服务器将所请求的对象放在 HTTP 响应报文中返回给校园网的高速缓存。



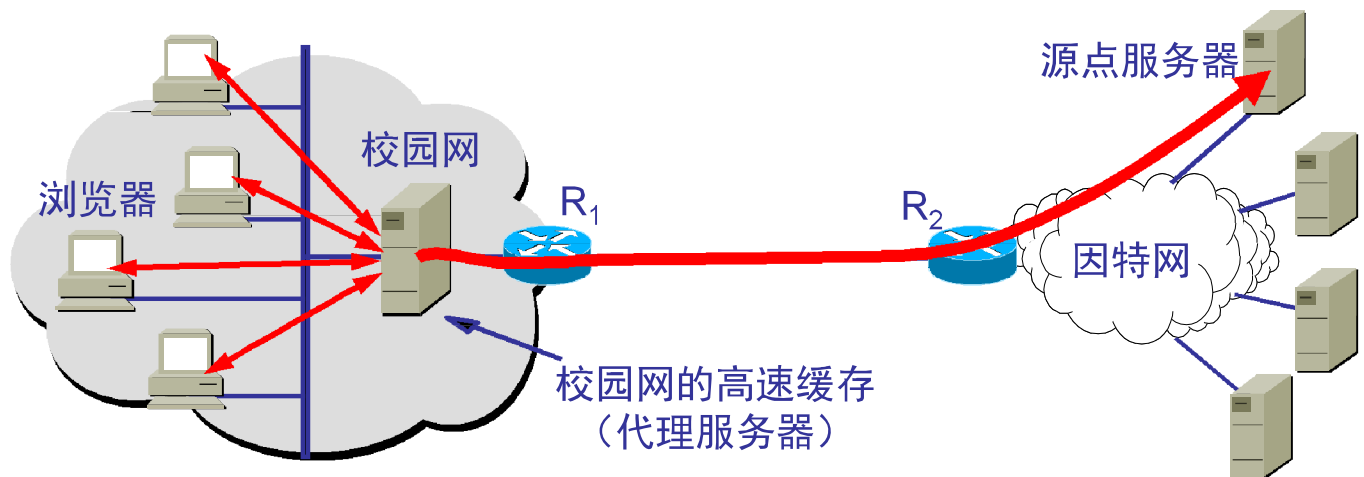
使用高速缓存的情况

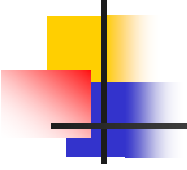
(5) 高速缓存收到此对象后，先复制在其本地存储器中（为今后使用），然后再将该对象放在 HTTP 响应报文中，通过已建立的 TCP 连接，返回给请求该对象的浏览器。



使用高速缓存的情况

(6) 代理服务器的另一个作用就是可以用来隔离内外网络。





6.3.4 万维网的文档

1. 超文本标记语言 HTML

- 超文本标记语言 HTML 中的 Markup 的意思就是“设置标记”。
- HTML 定义了许多用于排版的命令（即标签）。
- HTML 把各种标签嵌入到万维网的页面中。这样就构成了所谓的 HTML 文档。HTML 文档是一种可以用任何文本编辑器创建的 ASCII 码文件。



HTML 文档

- 仅当 HTML 文档是以.html 或 .htm 为后缀时，浏览器才对此文档的各种标签进行解释。
- 如 HTML 文档改换以 .txt 为其后缀，则 HTML 解释程序就不对标签进行解释，而浏览器只能看见原来的文本文件。
- 当浏览器从服务器读取 HTML 文档后，就按照 HTML 文档中的各种标签，根据浏览器所使用的显示器的尺寸和分辨率大小，重新进行排版并恢复出所读取的页面。



HTML 文档

- HTML允许在万维网页面中插入图像。
- HTML还规定了链接的设置方法。
- 链接的终点可以是其他网站上的页面。这种链接方式叫做**远程链接**。
- 有时链接可以指向本计算机中的某一个文件或本文件中的某处。这叫做**本地链接**。



HTML 文档

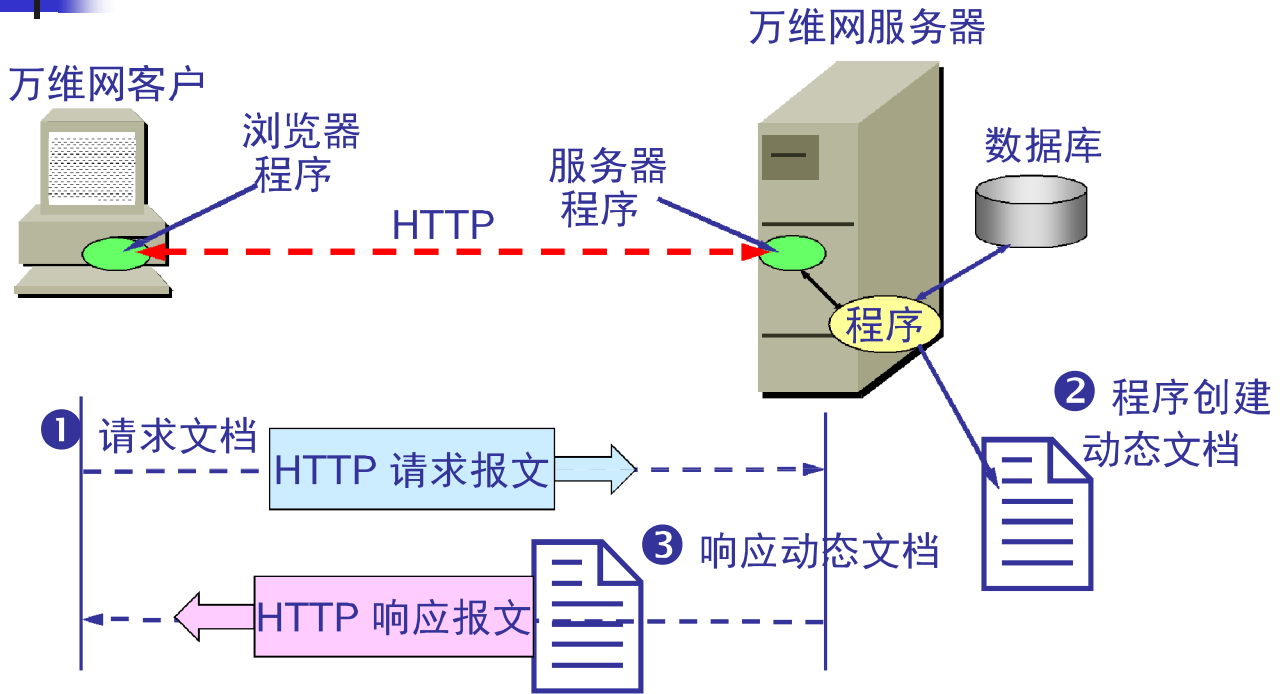
- 虽然完全可以用任何文本编辑器来编辑 HTML 文档，但使用“**所见即所得**”的万维网页面开发工具能很方便地制作各种美观的页面。
- 目前较为流行的网页制作工具有 FrontPage, DreamWeaver 等。



2. 动态文档

- **静态文档**是指该文档创作完毕后就存放在万维网服务器中，在被用户浏览的过程中，内容不会改变。
- **动态文档**是指文档的内容是在浏览器访问万维网服务器时才由应用程序动态创建。
- 动态文档和静态文档之间的主要差别体现在**服务器**一端。这主要是文档内容的生成方法不同。而从浏览器的角度看，这两种文档并没有区别。

动态文档





动态文档技术

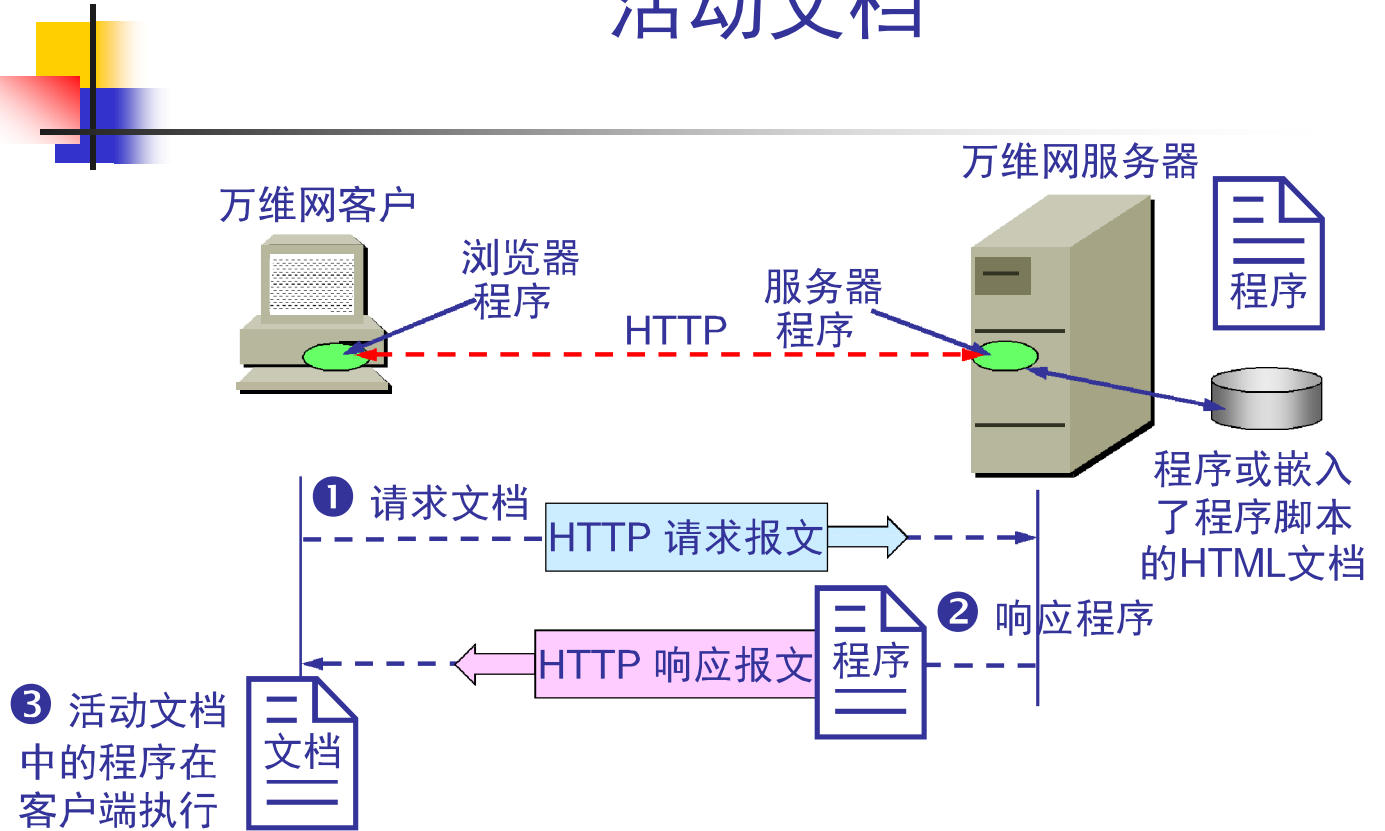
- 通用网关接口(Common Gateway Interface, **CGI**)
- 超文本预处理器(Hypertext Preprocessor, **PHP**), 使用Perl语言
- Java服务器网页(Java Server Pages, **JSP**), 使用Java语言
- 活动服务器网页(Active Server Pages, **ASP**), 使用VBScript, JScript等语言
- ASP.NET, 使用C#, VB.net等语言



3. 活动文档

- **活动文档**(active document)技术把所有的工作都转移给浏览器端。
- 每当浏览器请求一个活动文档时，服务器就返回一段程序副本在浏览器端运行。
- 活动文档程序可与用户直接交互，并可连续地改变屏幕的显示。
- 由于活动文档技术不需要服务器的连续更新传送，对网络带宽的要求也不会太高。

活动文档





活动文档技术

- Java applet
- JavaScript
- ActionScript
- 等等



混合文档

- 实际上，现在万维网上的很多文档都是这三种文档的混合体。
- 在这样的万维网页面中有一部分是用HTML编写的静态部分，一部分是用程序在服务器端动态生成的，还有一部分是可以在浏览器端运行的程序或程序脚本。



4. B/S应用程序结构

- 浏览器/服务器(Browser/Server)方式，一种特殊的C/S方式
- 利用动态和活动网页，通过通用的浏览器为用户提供人机交互的界面
- 优点是用户不需要安装单独的应用程序，简化了应用的开发、维护和使用
- 越来越多的网络应用采用B/S结构，例如购物网站、电子邮件、搜索引擎、博客等等。



6.3.5 移动Web

- 早期采用新的协议栈：无线应用协议 WAP (Wireless Application Protocol)，但随着网络带宽和设备计算能力的提高，目前更多的方法是：

6.3.5 移动Web

- 1. **移动版本网页**。现在越来越多的Web网站针对移动电话用户开发和设计移动友好(mobile-friendly)的Web页面内容。
- 2. **内容转换技术**。利用设置在移动电话和Web服务器之间的转码服务器，将页面内容转换成移动友好的内容再发送给用户。
- 3. **移动浏览器**。为手持设备的小型屏幕显示网页做了各种优化，通常与转码器技术配合使用，以减少产生的流量。例如全球著名的Opera，国内的UC浏览器。



6.3.6 万维网搜索引擎

- 搜索引擎实际上就是一个基于B/S结构的网络应用软件系统
- 从网络用户角度来看，它根据用户提交的类自然语言查询词或者短语，返回一系列很可能与该查询相关的网页信息，供用户进一步判断和选取。
- 搜索引擎要尽量提高**响应时间**、**查全率**、**查准率**和**用户满意度**四个指标，即用尽可能少的时间返回尽可能相关的网页信息列表，并将最可能满足用户需求的信息排在最前面。



(1) 网页搜集

- 大规模搜索引擎都是事先通过网页搜集软件在万维网上自动搜集大量网页并下载存储在本地存储系统中供以后进行查询。
- 通过网页之间的超链关系，网页搜集软件按照先深或先广遍历算法在万维网上从一个网页查找到另一个网页，就好像蜘蛛在蜘蛛网上爬行一样，因此网页收集软件往往被称为“蜘蛛”或“**网络爬虫**”。



(2) 建立索引

- 针对每个查询请求直接到这些海量网页中去全文检索太慢。
- 更好的方法是事先遍历每个网页并记录每个网页包含的各种词汇及其位置，然后根据词汇反过来建立索引项记录包含该词汇的网页及位置。以后根据关键词检索网页就非常迅速了。
- 将文档以关键词作为索引的数据结构被称为**倒排表**，是进行快速全文检索的关键。



(3) 检索排序

- 搜索引擎需要将与用户查询条件相关性最高的网页条目放在最前面。
- 最简单的方法就是根据查询关键词在网页上出现的频率进行排序。
- Google综合考虑了网页的**重要性**和**相关性**，认为被链接得越多的网页越重要，被越重要的网页链接的网页也越重要，其**PangeRank算法**可以快速地计算每个网页的重要性排名。Google因此成为了最优秀的搜索引擎。



2. 垂直搜索引擎和元搜索引擎

- **垂直搜索引擎**(Vertical Search Engine), 它们针对某一特定领域、特定人群或某一特定需求提供搜索服务。目前热门的垂直搜索领域有：购物、旅游、汽车、论坛、房产、求职、交友、图片等等。
- **元搜索引擎**(Meta Search Engine)是搜索引擎之上的搜索引擎，它们自己并不在万维网上搜集网页，而是在接受用户查询请求时，同时在其他多个搜索引擎上进行搜索，并将检索的结果进行综合处理后，以统一的格式返回给用户。



6.3.7 博客与微博

1. 博客

- **博客**是万维网日志(web log)的简称。也有人把blog进行音译,译为“部落格”,或“部落阁”。还有人用“博文”来表示博客文章。
- 在博客出现以前,网民是因特网上内容的消费者。但博客改变了这种情况,网民不仅是因特网上内容的消费者,而且还是因特网上内容的生产者。
- 现在从一些著名的门户网站的主页上都可以很容易地进入到博客的页面,这让用户查看或发表自己的博客都是非常方便的。



博客与个人网站的区别

- 建立个人网站不仅的成本较高，需要租用个人空间、域名等，同时对建立网站的个人需要懂得HTML语言和网页制作等相关技术
- 博客在这方面是不需要什么投资的，所需的技术仅仅是会上网和会用键盘或书写板输入汉字即可。
- 因此网民用较短的时间就能够把自己写的博客发表在网上，而不像制作个人网站那样花费较多的时间。



2. 微博

- 微博不同于一般的博客。微博只记录片段、碎语，三言两语，现场记录，发发感慨，晒晒心情，永远只针对一个问题进行回答。
- 根据新浪微博白皮书，从2010年3月到2010年6月，新浪微博月覆盖人数从2510.9万增长到4435.8万。
- 博客或微博里的朋友，常称为“博友”。微博也被人戏称为“围脖”，因此现在也有人把博友戏称为“脖友”。